Identifying predictors of physics item difficulty: A linear regression approach

Vanes Mesic and Hasnija Muratovic

Faculty of Science, University of Sarajevo, Zmaja od Bosne 35, 71000 Sarajevo, Bosnia and Herzegovina

(Received 30 October 2010; published 10 June 2011)

Large-scale assessments of student achievement in physics are often approached with an intention to discriminate students based on the attained level of their physics competencies. Therefore, for purposes of test design, it is important that items display an acceptable discriminatory behavior. To that end, it is recommended to avoid extraordinary difficult and very easy items. Knowing the factors that influence physics item difficulty makes it possible to model the item difficulty even before the first pilot study is conducted. Thus, by identifying predictors of physics item difficulty, we can improve the test-design process. Furthermore, we get additional qualitative feedback regarding the basic aspects of student cognitive achievement in physics that are directly responsible for the obtained, quantitative test results. In this study, we conducted a secondary analysis of data that came from two large-scale assessments of student physics achievement at the end of compulsory education in Bosnia and Herzegovina. Foremost, we explored the concept of “physics competence” and performed a content analysis of 123 physics items that were included within the above-mentioned assessments. Thereafter, an item database was created. Items were described by variables which reflect some basic cognitive aspects of physics competence. For each of the assessments, Rasch item difficulties were calculated in separate analyses. In order to make the item difficulties from different assessments comparable, a virtual test equating procedure had to be implemented. Finally, a regression model of physics item difficulty was created. It has been shown that 61.2% of item difficulty variance can be explained by factors which reflect the automaticity, complexity, and modality of the knowledge structure that is relevant for generating the most probable correct solution, as well as by the divergence of required thinking and interference effects between intuitive and formal physics knowledge structures. Identified predictors point out the fundamental cognitive dimensions of student physics achievement at the end of compulsory education in Bosnia and Herzegovina, whose level of development influenced the test results within the conducted assessments.

DOI: [10.1103/PhysRevSTPER.7.010110](https://doi.org/10.1103/PhysRevSTPER.7.010110)

PACS numbers: 01.40.Fk, 01.40.gf

I. INTRODUCTION

Physics education quality improvement can be achieved by developing a functional iterative cycle that consists of curriculum programming, instruction, and assessment. According to Redish [1], each of these fundamental elements should take into account a model of student cognitive and affective functioning. We cannot directly observe the cognitive and affective functioning of our students. Various aspects of student functioning can be realized only after having studied student behavior in concrete situations. The credibility of the developed student model grows with the number of different situations the student has encountered. The most practical way for confronting students with concrete physical situations is to administer a physics test to them. The higher the number and versatility of used items, in regards to tapping various aspects of physics competence, the higher the probability

of obtaining a more appropriate student model by analyzing the test results.

Quality management in physics education urges for feedback on student cognitive achievement that is based on testing representative student samples. Hence, it is important to conduct large-scale assessments of student achievement in physics, as well as to analyze and use the results of those assessments. Thus far, students from Bosnia and Herzegovina have participated in two large-scale assessments of cognitive achievement in physics. In 2006, the local Standards and Assessment Agency (SAA) conducted a large-scale study of cognitive achievement in physics at the end of compulsory education (eighth or ninth grade students, depending on region) in Bosnia and Herzegovina. This study was based on local curricula existing at that time, but no explicit assessment frameworks were created, which made it difficult to impute a qualitative meaning to quantitative test results [2]. Moreover, within conducted pilot studies a significant number of created items displayed poor psychometric characteristics and had to be discarded. In most cases the low discriminatory power of those items was related to their high difficulty [2]. One year after the first large-scale

Published by the American Physical Society under the terms of the Creative Commons Attribution 3.0 License. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

assessment of physics achievement, students from Bosnia and Herzegovina participated in the Trends in International Mathematics and Science Study (TIMSS). TIMSS has been conducted in four-year cycles. It incorporates assessments of student mathematics and science achievement at the end of fourth and eighth grade, as well as collecting data about teaching and learning contexts in each participating country. Within TIMSS assessment frameworks physics content areas and categories of cognitive activities are specified [3]. Each physics item is assigned to only one cognitive category and one physics content area. Such a practice of a universally relevant classification of items is highly questionable—students from countries where certain physical phenomena are to be explicitly elaborated in physics instruction could solve the corresponding items by rote memorization, whereas students from other countries would have to be engaged in higher thinking processes.

Primary analysis of the data obtained within the above-mentioned assessments pointed out the low values of quantitative achievement measures [2,4], but it remained unclear which achievement factors gave rise to such results. In order to receive useful feedback for all the participants of the physics education process at the level of compulsory education in Bosnia and Herzegovina, we attempted to identify the factors which had made the physics items more or less difficult for students from Bosnia and Herzegovina, as well as to rank them with respect to their importance.

In addition to feedback on curriculum implementation, the practical importance of this study is reflected in the potential improvement of the test-design process. According to Chalifour and Powers [5], “besides needing to meet specifications for content, test developers must also generate items having appropriate degrees of difficulty.” The item difficulty can be known only after piloting the test [6], whereby, based on item response theory (IRT) analysis, items with poor psychometric features are often automatically discarded. Therefore, the number of test items that must be developed is sometimes much greater than the number that is eventually judged suitable for use in operational test forms [5]. Rosca [7] points out that IRT models do not specify the item characteristics which make some items more or less difficult for students and that “information regarding what factors impact the item difficulty can be used by test developers to wield some control over the item difficulty of the items included in a test.” Taking into account the presented references, we believe that the method presented in this study could help test developers to reduce the size of the initial item pool required by large-scale studies. Instead of discarding interesting test items with poor psychometric characteristics in preliminary IRT analysis, test designers could systematically modify them with information obtained from linear regression analysis of item difficulty. The same information could also be used for designing items of various difficulties to assess fundamental aspects of physics competencies.

The theoretical significance of the study that is presented in this paper is reflected in determining some relatively independent cognitive dimensions of physics competence. In other words, we expect to gain additional insight into the structure of physics competence by evaluating and categorizing the identified predictors of item difficulty.

II. REVIEW OF THE LITERATURE

Within the relevant scientific literature on item difficulty issues, the linear regression approach is predominantly used.

Rosca [7] conducted a study with the purpose of identifying factors that made the TIMSS 2003 science items difficult. Based on her study of relevant literature, she singled out 17 potential predictors of item difficulty. Those predictors were related to item textual properties, the elicited cognitive demand, the corresponding science domain, and response selection properties. Thereafter, Rosca performed an item analysis with respect to singled-out potential predictors and calculated Rasch item difficulties for the U.S. student sample. For this purpose, she used 104 multiple-choice items from the TIMSS 2003 science assessment. Statistical significance and relative importance of potential predictors was tested by creating a regression model of item difficulty. The created model made it possible to explain 29.8% of item difficulty variance by means of Flesch reading ease score, ratio of the number of words in the solution and average number of words in distractors, cognitive level according to Bloom, average number of words in distractors, and the presence of graphics in the item stem. All predictors, besides the Flesch reading ease, were significant at the $p < 0.1$ level, and most of the explained variance could be assigned to the predictor “cognitive level according to Bloom.”

According to Weinert [8], competencies represent “to individuals available or by means of them accessible skills and abilities which are used for problem solving, as well as related motivational, conative and social aptitudes and skills which make it possible to readily and efficiently utilize the problems’ solutions in variable situations.” By performing a logical analysis of physics competence, Kauertz [9] came to the conclusion that it could be modeled based on combinations of cognitive activities, content complexity, and guiding ideas.

Guiding ideas are supposed to be basic physics concepts or formalisms that can be a starting point for effective structuring of physics contents (e.g., concepts of energy, interaction, systems and matter, mathematical formalism, etc.). Regarding the cognitive activities dimension of physics competence, Kauertz differentiates between processes of “knowing,” “structuring,” and “exploring.” Thereby, structuring refers to organizing of the existing knowledge base, whereas exploring includes discovering new relationships. Kauertz’s content complexity can be described by

six hierarchically arranged levels: one fact (I), several facts (II), one relationship (III), several unrelated relationships (IV), several related relationships (V), basic concept (VI).

Starting from the physics competence model, as described above, Kauertz [9] created 120 physics items and conducted a study, whereas the student sample consisted of 535 10th grade students from Germany. Then, he ran a factorial analysis of variance (ANOVA) of item difficulty whereas the factors were the physics competency dimensions, as well as the interactions “complexity and guiding idea” and “complexity and cognitive activity.” Thus, 52.4% of item difficulty variance could be explained, but the model as a whole was not statistically significant. Only “content complexity” and “guiding idea” proved to be statistically significant factors. The corrected model accounted for 23.7% of item difficulty variance, whereas a much bigger effect was reported for content complexity than for the guiding idea factor.

Hotiu [10] studied the relationship between item difficulty and item discriminatory power for purposes of improving the test-design process within the physical science course at Florida Atlantic University.

She developed a method for assigning difficulty levels to multiple-choice items. By accommodating Bloom’s taxonomy, she ranked the difficulty levels of activities that are relevant for solving physics items (see Table I). Then she calculated the overall item difficulty level by adding up the difficulty levels of all the activities that one has to implement when solving that item.

Hotiu came to the conclusion that items with a difficulty level between 9 and 14 display the best discriminatory behavior (discriminatory index above 0.6).

Considering the results from the conducted studies, we can conclude that a rather big part of item difficulty variance could not be accounted for by the mentioned predictors. We can assert that the relevant results of physics education research relating to student cognitive functioning issues have not been taken into consideration sufficiently. Interference effects between intuitive and formal

physics knowledge structures has not been addressed in any of the described studies, as well as the importance of divergent thinking. Only Hotiu specified some factors which partly describe the ability of using various representations of physics knowledge. In addition, it is clearly established that most of the predictors that reflect items’ formal features cannot account for bigger portions of item difficulty variance.

III. MATERIALS AND METHODS

A. Student sample

In 2006, SAA conducted an assessment of student achievement in physics at the end of compulsory education in Bosnia and Herzegovina. 1377 students participated in that study. One year later, 4220 students of same age as in the previous study (mostly 14 year old) participated in TIMSS. In both studies, the student sample was generated by stratified sampling of students from entire Bosnia and Herzegovina [2,4]. The student samples were representative.

B. Item sample

According to science item almanacs [11] the TIMSS 2007 test booklets included 59 physics items, whereas the SAA 2006 test booklets included 64 physics items. Within the whole sample of 123 physics items, there were 66 multiple-choice items and 57 constructed-response items. In both studies, the students did not have to solve all of the physics items because a matrix test design and IRT test scoring were used [2,4]. Each of the TIMSS items was administered to approximately 600 students, and each of the SAA physics items was administered to approximately 450 students. The TIMSS 2007 physics items were created along the lines of TIMSS assessment frameworks, and the SAA assessment of physics achievement was based on the local curricula that were current in 2006. Within the SAA study no explicit assessment framework was used.

C. Design and procedures

Taking into account that the physics item difficulty significantly depends on certain cognitive aspects of students’ physics competencies, we studied the relevant literature with the purpose of identifying constructs that define the cognitive dimension of physics competence. Thereafter, we performed an item content analysis with respect to the identified cognitive constructs as variables. Mostly, these cognitive constructs were characterized by a hierarchical structure, so we had to describe items by multiple level variables.

Each item was associated with only one level of each variable. When we were classifying items with respect to the allocated types of knowledge or cognitive processes, we assigned the item to the highest allocated level of the

TABLE I. Classification of performance tasks by means of difficulty level.

Difficulty level	Performance tasks
1	Knowledge and remembering
2	Identifying
3	Applying
4	Simple unit conversion
5	Simple equation
6	Unit conversion
	Vector analysis
5	Solving an equation
	Derivation
6	Solving systems of equations

correspondent variable within the most probable solution [12]. In the case of several variables the variable levels were created in an empirical manner by implementing processes of item differentiating with respect to the correspondent cognitive construct.

In order to perform quantitative item analysis, we created an item database by using the SPSS software. The database contained information regarding the 123 physics items from the conducted large-scale assessments. We described items only by those variables (see Table II) whose levels could be associated with at least 10 items. Because of an insufficient number of physics items that could be associated with the processes of analogical and extreme case reasoning, we had to discard these potential predictors, albeit they were supposed to be very important

for physics [18,20–22]. For some variables the problem has been solved by collapsing similar variable levels, so that in the end a sufficient number of items was associated to each of the variable levels. Thus, for the original Kauertz content complexity variable, we collapsed the levels “one relationship” and “several unrelated relationships” (we obtained the level “relationships”), as well as the levels “several related relationships” and “basic concept” (we obtained the level “related relationships”). Finally, the levels “one fact” and “several facts” were collapsed to obtain the level “declarative knowledge”. Thus, the variable “modified Kauertz content complexity” has been created. Its baseline category (declarative knowledge) can be used to describe items which require static knowledge, whereby the other two levels (relationships and

TABLE II. Potential predictors of item difficulty.

Variable name	Levels of the variable	Reference
Modified Kauertz content complexity	0—declarative knowledge 1—relationships (including rules of their use) 2—related relationships (including the rules of their use)	[9]
Analytic content representation	0—does not require the use of analytic representation 1—requires the use of analytic representation	[10]
Knowledge of experimental method	0—does not require knowledge of experimental method 1—requires knowledge of experimental method	Personal experience
Interference effects of intuitive and formal physics	0—negligible interference effects 1—intuitive thinking facilitates item solving 2—counterintuitive thinking is necessary for item solving	[13–15]
Cognitive activities	0—remembering 1—“near” transfer 2—exploration	[9]
Divergent thinking	0—does not require divergent thinking 1—requires divergent thinking	[16]
Visualization	0—visualization is not important for item solving 1—visualization is important for item solving	[17,18]
Mitigating factors	0—there are no mitigating factors for item solving 1—item can be solved by remembering little fragments of knowledge (symbols of physical units and quantities, often used graphical symbols), or by remembering fundamental physical laws or formulas that are explicitly used in a great number of occasions, or if the item can be solved without the use of formal physics knowledge	Content analysis of empirically easiest physics items; collapsing of several variables
Item openness	0—multiple-choice items (4 options) 1—constructed-response items	[19]
Presence of graphics in the item stem	0—item stem does not contain graphics 1—item stem contains graphics	[7]
Number of words in item stem	Continuous variable	[7]

TABLE III. Percent of items through characteristic intervals of out-fit and in-fit values.

Out-fit	0.5–0.7	0.71–0.85	0.86–1.15	1.16–1.30	1.31–1.50
TIMSS	1.7%	0%	94.9%	1.7%	1.7%
SAA	9.4%	6.3%	79.7%	3.1%	1.6%
In-fit	0.5–0.7	0.71–0.85	0.86–1.15	1.16–1.30	1.31–1.50
TIMSS	0%	1.7%	98.3%	0%	0%
SAA	0%	0%	100%	0%	0%

related relationships) can be used to describe the complexity of schematic knowledge required by some items. Thereby, the “schematic knowledge” construct represents “knowledge which combines procedural and declarative knowledge” [23]. The “mitigating factors” variable was mostly created by collapsing the “fragments of knowledge” variable, obtained by content analysis of empirically easiest physics items, with extreme levels of the “positive influence of intuitive physics” variable. Actually, by using processes of comparing and differentiating items which (most probably) activate intuitive physics knowledge, we could distinguish between items which can be (most probably) solved without any prior formal physics education and items for which the intuitive physics could only facilitate item solving, but they still require some formal physics education. All items for which we judged to be coded by one for the mitigating factors variable share a common feature—the answer to them is most probably highly automated.

With the purpose of evaluating the importance and statistical significance of singled-out potential predictors, we had to establish a relationship between these theoretical item descriptors and an empirical measure of item difficulty. Therefore, we decided to calculate the Rasch item difficulties for all 123 included physics items. Taking into account that the focus of our study was on item difficulty rather than on other parameters, we chose to use the Rasch simple logistic model. For this purpose, it was necessary to recode student answers from primary student achievement databases [11,24]. Because we decided to use the one parameter model, all the partially correct answers had to be considered as incorrect. The correct answers were coded by 1, and incorrect answers by 0. Thereafter, the student achievement data were stored in two separate text files (one for each of the large-scale assessments) where rows of data represented individual students and columns of data represented individual items. Based on the student achievement data that were given in these text files, the Acer CONQUEST 2.0 software [25] generated, in separate analyses, estimations of item difficulties and correspondent item fit statistics (see Table III).

Items which are sufficiently in accordance with the Rasch model to be productive for measurement have in-fit and out-fit values between 0.5 and 1.5 [26,27]. Thus, by inspecting Table III, we could conclude that the goodness of fit for items which were used in our study is satisfying.

Further, to make the item difficulties from two different assessments comparable, a virtual test equating procedure had to be implemented [28]. This technique of test equating is to be used in circumstances where both the student sample and the item sample are different for two assessments (there are no “common” students or “common” items), but the items cover similar material [28,29]. The steps of the virtual test equating procedure are as follows:

(1) Identifying pairs of items (one from each study) that are as close as possible similar to each other, with respect to physics content and estimated difficulty. It is necessary to have at least five pairs of items. In this study, we chose 10% of the questions, which is six pairs, as the basis of equating.

(2) Cross-plotting the corresponding item difficulties, with item difficulties from the more reliable assessment represented on the x axis.

(3) Fit the data in step (2) with a linear line.

(4) Rescaling of item difficulties for the assessment that was represented on the y axis of the item difficulty cross-plot. It is necessary to multiply each of these item difficulties with the reciprocal slope value and to add the x -intercept value of the fit line to the result of the performed multiplication: $\text{TEST } Y_{\text{TEST } X\text{-frame}} = \text{TEST } Y_0 \times \frac{1}{k} + (-\frac{a}{k})$.

The cross-plot of item difficulties that was created for the purposes of this study is given in Fig. 1.

Based on the fit line slope and x -intercept value, we rescaled the item difficulties for the SAA assessment. Therefore, at the end we could assign to all 123 physics items empirical difficulty measures and all of those measures were comparable.

Now, it was possible to quantify the statistical significance and relative importance of the singled-out potential item difficulty predictors. For this purpose, we decided to create a linear regression model of physics item difficulty. First, we had to check if the size of our item sample was big enough for regression analysis purposes. According to Miles and Shelvin [30], if we expect to obtain a large effect, it is sufficient to have 80 items of analysis. Clearly, this condition has been met.

Further, for categorical variables with more than two levels, a dummy-coding procedure had to be implemented [31]. There were three variables with more than two categories (see Table II)—modified Kauertz content complexity, cognitive activities, and interference effects of intuitive

FIG. 1. Cross-plot of item difficulties for six item pairs from our study.

and formal physics. Thereby, for these three variables, we chose declarative knowledge, remembering, and negligible interference effects to represent baseline categories, respectively. Out of the remaining levels of the mentioned variables, six potential predictors were obtained: relationships, related relationships, near transfer, exploration, positive influence of intuitive physics, and negative influence of intuitive physics.

After the dummy-coding had been done, we ran the linear regression procedure within SPSS 17.0. Thereby, the backward method was selected because we had no insight into the relative importance of the singled-out potential predictors of item difficulty. Within this method all potential predictors are entered into the initial model and the software sorts out only statistically significant predictors [31]. Statistically significant predictors which were identified by means of the described method constitute the final model of physics item difficulty (see Table VI).

Finally, we assessed the obtained model. For this purpose, we first examined if there were outliers or influential cases. Then we checked the linear regression assumptions. Field [31] suggests to always check the assumptions of independence and normal distribution of the residuals, as well as the linearity and homoscedasticity assumptions.

The functionality of the created model depends on the reliability of item analysis with respect to the identified predictors of item difficulty. For the purposes of checking

the interrater reliability, an item coding instruction was created (see Appendix B). Then we selected two postgraduate students with experience of work in school and organized a short item coding training for them. First, the coders were instructed about some prominent characteristics of identified item difficulty predictors. Then we selected three physics items out of our item sample and demonstrated how to use the item coding instruction. Afterwards the coders analyzed four additional items in a “think-aloud” manner, and we discussed the problems they had encountered while coding these items. Finally, the coders were asked to perform coding of 40 released physics items from the conducted assessments. We used Fleiss’ kappa [32] as a measure of intercoder agreement because there were more than two coders—the first author of this paper and two postgraduate students.

IV. RESULTS

A. Basic features of the obtained item difficulty model

The following potential predictors were entered into the initial model: “analytic representation,” “mitigating factors,” “experimental method,” “item openness,” “relationships,” “related relationships,” “positive influence of intuitive physics,” “negative influence of intuitive physics,” “near transfer,” “exploration,” “number of words in item stem,” “presence of graphics in item stem,” “visualization,” and “divergent thinking.”

The implementation of the backward method upon this set of potential predictors finally gave rise to a model of physics item difficulty whose basic features are given in Table IV.

The obtained model makes it possible to explain 61.2% of item difficulty variance. A rather small difference between R^2 and adjusted R^2 indicates the possibility of model generalization. Only item difficulty predictors that proved to be statistically significant at the $p < 0.05$ level remained in the model—labels of correspondent variables are specified below Table IV.

Results of the ANOVA procedure are given in Table V.

We can conclude that the regression model as a whole is statistically significant—the probability of obtaining such a large F -statistics value by chance is less than 0.1%.

Table VI provides information on some prominent features of item difficulty predictors that proved to be statistically significant.

TABLE IV. Model^a summary.

R	R square	Adjusted R square	Std. error of the estimate	Durbin-Watson
0.782	0.612	0.588	0.730 790	1.846

^aPredictors: (Constant), analytic representation, mitigating factors, experimental method, relationships, positive influence of intuitive physics, item openness, related relationships. Dependent variable: Rasch item difficulty.

TABLE V. ANOVA.

	Sum of squares	d.o.f.	Mean square	<i>F</i>	Sig.
Regression	96.850	7	13.836	25.907	0.000
Residual	61.416	115	0.534		
Total	158.266	122			

Based on standardized β coefficients we can rank statistically significant predictors with respect to the size of their unique influence on item difficulty. The predictor analytic representation exerts the largest influence on item difficulty followed by mitigating factors, item openness, related relationships, positive influence of intuitive physics, relationships, and experimental method.

Thus far, we have pointed out the factors that influence physics item difficulty and compared them with respect to

their relative importance. For the purposes of getting some more feedback on physics education at the primary school level in Bosnia and Herzegovina, it is useful to analyze an additional, absolute measure of students' physics achievement. Therefore, we decided to calculate classical item difficulties for categories of items which are described by the identified predictors of item difficulty (see Table VII).

B. Identification of potential outliers and influential items

By performing casewise diagnostics, we identified six outliers (see Table VIII).

The proportion of items whose standardized residuals are above 2 is below 5%, and the proportion of those items whose standardized residuals are above 2.5 is less than 1%. These values are tolerable [31].

TABLE VI. Predictor statistics.

Predictor	<i>B</i>	Std. error	Beta	<i>t</i>	Sig.	Tolerance
(Constant)	−0.209	0.148		−1.410	0.161	
Item openness	0.639	0.144	0.281	4.456	0.000	0.848
Positive influence of intuitive physics	−0.581	0.181	−0.206	−3.211	0.002	0.820
Relationships	0.334	0.162	0.142	2.060	0.042	0.713
Related relationships	0.691	0.187	0.267	3.689	0.000	0.644
Experimental method	0.609	0.275	0.140	2.209	0.029	0.844
Mitigating factors	−0.811	0.175	−0.292	−4.622	0.000	0.846
Analytic representation	0.993	0.202	0.309	4.903	0.000	0.848

TABLE VII. Percent of correct answers with respect to categories of statistically significant predictors; coding is in line with the item coding instruction (see Table XII).

Item openness		Mitigating factors		Analytic representation		Experimental method	
0	1	0	1	0	1	0	1
42.47	26.00	29.4	55.13	37.78	17.69	35.55	25.83
Intuition (positive)		Relationships		Related relationships			
0	1	0	1	0	1		
31.93	46.22	37.00	31.08	39.22	22.38		

TABLE VIII. Casewise diagnostics^a.

Case number	Std. residual	Rasch difficulty	Predicted value	Residual
59	−2.075	0.240	1.756 51	−1.516 507
70	−2.109	−1.111	0.430 23	−1.541 173
86	2.673	2.717	0.763 96	1.953 174
88	2.190	3.714	2.113 48	1.600 325
97	2.474	2.572	0.763 96	1.807 825
119	2.283	3.782	2.113 48	1.668 432

^aDependent variable: Rasch item difficulty.

FIG. 2. Cook's distances for used items.

By calculating Cook’s distances, we checked if there were any items that had exerted large influence on the model as a whole. According to Cook and Weisberg [33], values greater than 1 may be cause for concern. For all used items Cook’s distances were considerably below 1 (see Fig. 2).

For the purpose of measuring the influence of each item on the individual predictors, difference in beta (DFBeta) values for each predictor were calculated. These measures represent differences between β coefficients when one item is included and not included, respectively [31]. The largest DFBeta value is associated with the pair “item S042238B-knowledge of experimental method” and it amounts to 0.557. It is supposed that the standardized DFBeta should not be above 1 [31]. Clearly, this condition is met for the obtained model.

Thus, we can conclude that there were no influential items and that the model is stable.

C. Testing assumptions

1. Assumptions of independent residuals and absence of multicollinearity

In order to check the assumption of independent residuals, we calculated the Durbin-Watson statistics which tests for serial correlation between errors [31]. Values above 3 or below 1 indicate that this assumption is not met, and the value 2 is ideal [31]. For our model the value of Durbin-Watson statistics (see Table IV) is 1.846. This is close to the ideal value, so we can claim that the assumption of independent residuals has been met.

Based on the fact that the values of tolerance statistics (see Table VI) are significantly higher than 0.2 for all the item difficulty predictors, we can conclude that there is no multicollinearity between them.

TABLE IX. Normality checks for standardized residuals.

Kolmogorov-Smirnov			Shapiro-Wilk		
Statistic	d.o.f.	Sig.	Statistic	d.o.f.	Sig.
0.051	123	0.200 ^a	0.990	123	0.499

^aThis is a lower bound of the true significance.

TABLE X. White's test of no heteroskedasticity against heteroskedasticity of some unknown general form.

White's test statistics	Degrees of freedom ^a	$\eta_{df} (p = 0.05)$
34.69	24	36.42

^aFour dummy interactions proved to be constants and were automatically excluded from the model.

2. Assumption of normally distributed residuals

In order to check the assumption of normally distributed residuals, we calculated the Kolmogorov-Smirnov and Shapiro-Wilk statistics for standardized residuals (see Table IX). Generally, these tests compare scores in the sample to a normally distributed set of scores with the same mean and standard deviation [31].

Both of them proved to be not statistically significant. Thus, we can conclude that the distribution of standardized residuals does not significantly deviate from the normal distribution.

The skewness and kurtosis z scores amount to 1.174 and 0.24, respectively. These values are not significant at the $p < 0.05$ level.

Based on all of the obtained results, we can conclude that the assumption of normally distributed residuals has been met.

3. Assumptions of linearity and homoscedasticity

Originally, the assumptions of linearity and homoscedasticity were checked by analyzing a “standardized residuals versus standardized predicted values” plot (see Appendix A). Thereby, we came to the conclusion that the linearity assumption has been met, but suspected slight deviation from homoscedasticity. Therefore, we decided to additionally test the homoscedasticity assumption by calculating the White test statistics [34] for our model.

White's test is a test of the null hypothesis of no heteroskedasticity against heteroskedasticity of some unknown general form. It follows chi-square distribution.

From Table X, we can conclude that the value of White statistics is lower than the correspondent value of chi-square statistics ($p = 0.05$). Thus the null hypothesis of homoscedasticity cannot be rejected.

TABLE XI. Intercoder agreement measures for singled-out item difficulty predictors.

	Item openness	Related relationships	Relationships	Mitigating factors
Fleiss' Kappa	1	0.67	0.62	0.64
	Experimental method	Positive influence of intuitive physics	Analytic representation	
Fleiss' Kappa	0.74	0.66	0.93	

D. Intercoder agreement

We calculated the interrater reliability measures for classifying items with respect to variables which proved to be statistically significant item difficulty predictors (see Table XI).

According to interpretation rules of kappa statistics, as were given by Landis and Koch [35], we can conclude that there was a substantial intercoder agreement for classifying items with respect to variables relationships, mitigating factors, positive influence of intuitive physics, related relationships, and experimental method. The intercoder agreement for item coding with respect to the variable analytic representation was almost perfect, whereas the classifying of items with respect to item openness was completely objective, as we had expected.

Fleiss [36] characterizes kappas of 0.60–0.75 as good and those over 0.75 as excellent

V. DISCUSSION

By creating the item difficulty model, we pointed out some of the basic ability factors that had influenced the physics item difficulty in a statistically significant manner.

The relative importance of singled-out item difficulty predictors can be assessed by comparing their standardized β coefficients [31].

Taking into account that Rasch difficulty is given in logits, and that “one logit is the distance along the line of the variable that increases the odds of observing the event specified in the measurement model by a factor of 2.718” [37], we will also discuss the influence of our predictors on odds of obtaining a correct answer.

Based on the comparison of standardized β coefficients for the predictors relationships and related relationships, we can conclude that the increasing of complexity of the knowledge structure, which is most probably used for item solving, causes the Rasch item difficulty to rise, provided that all other predictors are held constant. Thereby, if we increase the relationships and related relationships variables by one, the odds for obtaining a correct answer decrease by a factor of 1.39 and 2, respectively.

Taking into account that these variables reflect schematic knowledge, we also can come to the conclusion that items which tap schematic knowledge are significantly more difficult than items which tap declarative knowledge,

if we control the influence of the remaining variables from the model.

These conclusions are in line with the results of some previous studies [9]. According to de Jong and Ferguson-Hessler [38], one of defining features of declarative knowledge is its automaticity. In other words, such knowledge often can be processed automatically [39]. Actually, the influence of the knowledge complexity and automaticity factors on item difficulty can be partly explained by cognitive load theory [39]. In fact, the human short-term memory is very limited with respect to the number of elements (chunks) that can be held at the same time. Cognitive operations on these elements occupy additional space. Thus, clearly the cognitive demand increases with the number of activated relationships and with the need to perform operations on these relationships. It is very important to emphasize that the short-term memory is not limited with respect to the size of the chunks. Automated knowledge schemata induce negligible cognitive demand—one schema constitutes one chunk in the short-term memory [39].

According to results from Table VII, only one-third of students from Bosnia and Herzegovina succeeded to solve items that required the knowledge of relationships (including the rules of their use) and approximately one-fifth of them solved correctly items which required the knowledge of related relationships.

Taking into account the previously discussed statistically significant, unique effect of knowledge complexity and automaticity on item difficulty, as well as the very low student achievement on items that require schematic knowledge, we could conclude that the current physics instruction at the primary school level in Bosnia and Herzegovina mostly fails to foster students' schematic knowledge.

In that sense, it would be useful to pay more attention to developing an understanding of physical concepts and considering physics contents in various contexts, due to establishing strong and flexible links between physics concepts. It could be useful to reconsider the culture of setting and solving physics questions and problems in primary school physics education in Bosnia and Herzegovina. Thereby, questions or problems with higher intrinsic potential with respect to fostering conceptual knowledge should be preferred. The use of explicit conceptual maps

in physics instruction could also help students to build more functional knowledge structures.

“Knowledge of experimental method” proved to be a statistically significant predictor of item difficulty, too. The need for using the “knowledge of experimental method” causes an increase of Rasch item difficulty, provided that all the other predictors are held constant. Thereby, the odds of a correct response decreases by a factor of 1.84.

According to results from Table VII, approximately one-fourth of students from Bosnia and Herzegovina succeeded to solve items which required the knowledge of experimental method.

Taking into account the previously discussed statistically significant, unique effect of the knowledge of experimental method on item difficulty, as well as the very low student achievement on items that require experimental knowledge, we could conclude that the current physics instruction at the primary school level in Bosnia and Herzegovina mostly fails to foster the development of abilities related to planning, conducting, and analyzing experiments.

One of the main reasons for the low achievement of students from Bosnia and Herzegovina with respect to the knowledge of experimental method is the rare use of experimental method in schools in Bosnia and Herzegovina. In fact, according to results of TIMSS 2007, one-third of students from Bosnia and Herzegovina at the end of primary school education (eighth or ninth grade) claimed that they had never conducted a physics experiment on their own throughout their physics education [40].

With the purpose of improving the existing physics instruction practice in Bosnia and Herzegovina, prospective teachers should get into a habit of designing and conducting low-cost physics experiments. The knowledge of experimental method could be (partly) assessed by including appropriate items to written examinations, as it was done within TIMSS 2007.

Besides automaticity and complexity features of relevant knowledge schemes, the form of their representation affects the item solving efficacy, too. The standardized β coefficient for the predictor analytic representation is largest. In other words, in comparison to all the other predictors from the final model, the need for using the analytical representation has the largest impact on physics item difficulty. By increasing the analytic representation predictor by one, the Rasch item difficulty increases, if all the other predictors are held constant. Thereby, the odds of a correct answer decreases by a factor of 2.7.

Taking into account that 17 out of 18 items that required the use of analytic representation at the same time assessed the schematic knowledge of students and based on the statistical significance and sign of the analytic representation predictor, we can state that the item difficulty of items which assess schematic knowledge additionally increases if one has to use the analytic representation

of the relevant knowledge scheme in order to correctly solve the item, provided that all the other predictors are held constant.

According to results from Table VII, approximately 18% of students from Bosnia and Herzegovina succeeded to solve items which required the use of analytical representation.

Finally, we can conclude that the relatively low student performance on quantitative physics problems in the first place originates from students’ underdeveloped competencies of manipulating elements of schematic knowledge within the analytic form of representation.

The remaining of the positive influence of intuitive physics predictor within the item difficulty model once again confirms the importance of taking into account intuitive physics whenever we are to design physics classes. Rasch item difficulty decreases with an one unit increase of the positive influence of intuitive physics predictor, provided that all other predictors are held constant. Thereby, the odds of obtaining a correct answer increases by a factor of 1.79.

We should not only emphasize the negative aspects of intuitive physics, in the sense of physics misconceptions, but we should more often utilize its positive aspects for effectively building formal physics concepts [15].

“Mitigating factors” were mainly related to the need of remembering small fragments of knowledge or to the possibility of solving the item by utilizing given information without having to refer to physics knowledge. By increasing the mitigating factors variable by one the odds of a correct answer increases by a factor of 2.25, provided that all other predictors are held constant.

The statistical significance of this predictor is consistent with the significance of the knowledge complexity factor.

Within the set of predictors that reflect the items’ formal features, only the item openness predictor showed up to be statistically significant. The Rasch item difficulty increases if the students are required to construct a response by themselves, provided that all the other predictors are held constant. Thereby, the odds of obtaining a correct answer decreases by a factor of 1.89.

According to results from Table VII, for constructed-response items the average rate of students’ success has been 26%.

On the one hand, for multiple-choice items there is a possibility to solve the item correctly only by chance, and on the other hand, these items narrow the number of knowledge schemata that have to be evaluated in order to solve the problem. In other words, multiple-choice items possess a greater potential to guide students’ thoughts.

Regarding the predictors that proved to be nonsignificant at the $p < 0.05$ level, the largest partial correlation coefficients were associated with divergent thinking and counterintuitive thinking (see Table XIV). These predictors

were close to remaining in the regression model. One part of the item difficulty variance which was supposed to be explained by these predictors could be partly explained by some other predictors from the final regression model. Besides the fact that the divergent thinking predictor did not remain in the final item difficulty model, the importance of this cognitive construct is reflected in the statistical significance of item openness and experimental method predictors. In fact, by means of correlation analysis, it can be shown that divergent thinking correlates to the largest extent with these two predictors from the final model (see Table XIII). This correlation can be explained based on the asserted fact that multiple-choice items possess a “thought guiding” feature, as well as by taking into account the frequent need for designing subjectively new procedures in the case of items that elicit the knowledge of experimental method.

Surprisingly, the predictor counterintuitive thinking did not remain in the final model of item difficulty. This could be related to the fact that numerous quantitative items, for which the influence of intuitive physics was negligible, proved to be very difficult. The relatively small number of items that required counterintuitive thinking surely contributed to the nonsignificance of this predictor, too.

The predictor necessity of visualization proved to be nonsignificant. The largest part of item difficulty variance we supposed to be explained by this predictor could be explained by the predictor related relationships. The coefficient of correlation between these two predictors amounted to 0.509 (see Table XIII).

As well as in the study by Kauertz, cognitive activities proved to be nonsignificant at the $p < 0.05$ level. The use of more complex knowledge structures correlated with higher cognitive processes—the correlation coefficient between the variables “transfer” and “relationships,” as well as between the variables “exploration” and “related relationships,” has been above 0.7 (see Table XIII). Therefore, either knowledge qualities or cognitive processes could remain in the final model of item difficulty. Because of their higher partial correlation with item difficulty (see Table XIV), the knowledge descriptors remained in the model.

The predictors number of words in the stem and presence of graphics in the stem did not remain in the model of item difficulty. So, once again it has been shown that predictors that reflect the items’ formal features, with the exception of item openness, can account for only relatively small portions of item difficulty variance.

Based on the evaluation of the obtained results and on the categorization of the discussed cognitive constructs, it is possible to single out the following cognitive factor categories which influence the physics item difficulty:

(1) complexity and automaticity of knowledge structures which are relevant for generating the most possible solution,

(2) the predominantly used type of knowledge representation,

(3) nature of interference effects of relevant formal physics knowledge structures and correspondent intuitive physics knowledge structures (including p prims),

(4) width of the cognitive area that has to be “scanned” with the purpose of finding the correct solution and creativity,

(5) knowledge of scientific methods (especially experimental method).

According to the model of types and qualities of knowledge by de Jong and Ferguson-Hessler [38], automaticity, complexity, and modality come under fundamental qualities of knowledge. Thus, the structure of the obtained model of item difficulty is in line with the model of types and qualities of knowledge.

Besides general qualities of knowledge, our model also takes into account some cognitive domain features which are of particular interest for physics education (e.g., interference effects of intuitive and formal physics).

Regarding the model’s technical characteristics, we can say that the model as a whole is relatively stable and the linear regression assumptions are met.

The item coding interrater reliability is acceptable, but in the case of certain categories there is some place for improvement. Differences in intercoder agreement for coding the items with respect to different predictors emanate from differences in the nature of predictors, as well as from certain features of the item coding instruction. Thus, it is much easier to estimate if students had to use physical equations in order to solve one item than to estimate the probability of an item’s eliciting of intuitive physics knowledge or p prims. In fact, personal everyday experience, teaching experience, and theoretical knowledge on intuitive physics affect the coding of items with respect to the positive influence of intuitive physics predictor. Therefore, it could be useful to create lists of physics contents which most likely tap intuitive physics knowledge.

Regarding the coding of items with respect to types and qualities of knowledge, it has been shown that coders had more trouble with recognizing situations that require the use of one relationship than situations that require the knowledge of related relationships. In other words, for coders it was more difficult to estimate automaticity than complexity of knowledge.

For purposes of item coding with respect to the mitigating factors variable, it is necessary to define more precisely “physics knowledge elements which are explicitly stated and used in many occasions within physics education,” in order to improve interrater reliability.

Furthermore, it would be useful to specify additional criteria that would make it easier to decide whether or not one item, situated in the experimental context, can be solved without specialized knowledge of experimental method.

VI. SUMMARY AND CONCLUSION

The results of the study presented in this paper were obtained by combining qualitative and quantitative methods. Foremost, we studied the literature that is related to cognitive functioning of students with particular regard to physics learning and problem solving issues, as well as items that were used within large-scale assessments conducted in Bosnia and Herzegovina. Based on these analyses a list of potential predictors of item difficulty was created. After coding each of the items with respect to specified potential predictors, we calculated Rasch item difficulties. Finally, a regression model of item difficulty was created.

By means of the created model we can explain 61.2% of item difficulty variance. The structure of the obtained regression model is in accordance with the model of types and qualities of knowledge that was created by de Jong and Ferguson-Hessler [38]. Besides that, it takes into account some specific features of physics cognitive domain.

The functionality of the model is limited to samples of students at the end of compulsory education in Bosnia and Herzegovina, whereas the physics items should not elicit competencies that are fundamentally different from competencies we took into account within the process of model development.

We expect that the results of this study will underline the need for designing physics education at the primary school level in Bosnia and Herzegovina that will additionally take into account the importance of developing functional conceptual knowledge and basic knowledge of scientific methods, as well as the importance of intuitive physics and use of multiple representations of knowledge. In that sense, there is a need for change in the culture of setting and solving physics questions and problems. Namely, it is of particular interest to introduce new types of physics problems that possess a greater intrinsic potential to elicit higher cognitive processes.

Besides providing feedback for physics education at the primary school level in Bosnia and Herzegovina, the obtained model of physics item difficulty could improve the

test-design process for purposes of future large-scale assessments.

The obtained model reflects some fundamental aspects of physics competencies which cover most assessment goals of physics tests. When designing tests within the scope of TIMSS and SAA, the mathematical model of item difficulty in this study could also be used for manipulating the item features in the prepilot phase in order to get valid and psychometrically acceptable items. Such an approach would require detailed qualitative analysis of item features in the prepilot phase and their coding within correspondent statistical databases. The positive aspect of this method is also reflected in the fact that the process of assigning qualitative meaning to quantitative test results would become straightforward.

APPENDIX A: SUPPLEMENTARY AUXILIARY MATERIAL

See online supplementary material for further discussion regarding the fulfillment of the assumptions of linearity and homoscedasticity. Thereby, an alternative, graphical method to White's test is used which is based on analyzing certain characteristics of the standardized residuals versus standardized predicted values plot.

APPENDIX B: ITEM CODING INSTRUCTION

With the purpose of, as far as possible, reliable classification of items, it is necessary to have an idea of the prior physics knowledge of the tested target population. In that sense, for every individual item it can be recommended to investigate the ways in which the correspondent matter is instructed at school. To that end, it is possible to study the relevant physics curricula or standard physics textbooks used by the target population. The final coding of physics items should be based on the guidelines presented within the item coding instruction (see Table XII).

APPENDIX C: CORRELATION STATISTICS

The matrix of correlation coefficients (see Table XIII), for the potential item difficulty predictors and the

TABLE XII. Item coding instruction.

Variable	Levels of the variable	Indicators
Item openness	0—multiple-choice item 1—constructed-response item	Assign code 0 for multiple-choice items, and code 1 for constructed-response items.
Relationships	0—does not require knowledge of one or more unrelated relationships 1—requires knowledge of one or more unrelated relationships	Assign code 1 if it is necessary to use knowledge, but there is no need to establish links between several relationships or to consciously create links among various objects or situations, in general. An additional indicator for assigning code 1 is the opportunity to solve the item in only one step or in several, unrelated steps.

TABLE XII. (Continued)

Variable	Levels of the variable	Indicators
Related relationships	0—does not require knowledge of two or more related relationships 1—requires knowledge of two or more related relationships	Assign code 1 for all items that require combining two or more physical laws, that is in all cases where use of knowledge is required (negligible probability of giving an automatic response) and the item has not been encoded with 1 for the variable “relationships.” Also, assign code 1 if the student has to combine physics concepts in order to establish links between foreknowledge and concepts that were not explicitly stated within physics classes. In general, code 1 is assigned to items whose solution consists of several, interconnected steps.
Positive influence of intuitive physics	0—intuitive thinking facilitates item solving 1—intuitive thinking does not facilitate item solving	Assign code 1 if intuitive physics knowledge (knowledge on subjects of physical study, developed by means of everyday experience or “feeling” for physics phenomena) can significantly contribute to item solving. Encode in the same way items that are likely to elicit p primes, whereas these p primes positively contribute to item solving.
Analytic representation	0—use of analytic representation is not necessary 1—use of analytic representation is necessary	Assign code 1 if it asks for the use of the analytic representation of physical relationships (calculations based on physical formula, derivations, etc.).
Knowledge of experimental method	0—does not require knowledge of experimental method 1—requires knowledge of experimental method	Assign code 1 if students are required to rely on their knowledge of lab equipment or to think over experimental design. Use the same encoding if it is necessary to interpret a research experiment, whereas the student has to use specialized knowledge of experimental method in order to understand the experimental procedure. Assign code 0 if students are only asked to predict outcomes of simple demonstration experiments.
Mitigating factors	0—there are no mitigating factors 1—there are mitigating factors	Assign code 1 if the item can be solved by remembering small fragments of knowledge (symbols of quantities, units and prefixes; graphical symbols), as well as by solely remembering fundamental laws which are explicitly stated during physics lessons within a large number of teaching units. Apply the same encoding to items that can be solved without using formal physics knowledge, whereas the student does not have to use higher cognitive processes or intuitive physics knowledge.

empirical measure of item difficulty, provides valuable information about the interdependence of all cognitive constructs which were put into the initial regression model.

In order to draw conclusions about the unique influence of each potential item difficulty predictor, it is useful to analyze the correspondent coefficients of partial correlation (see Table XIV).

TABLE XIII. Zero-order correlation coefficients.

	Rasch difficulty	Item openness	Number of words	Divergent thinking	Presence of graphics	Intuitive physics (negative)	Intuitive physics (positive)	Near transfer	Exploration	Relationship	Related relationships	Visualization	Experimental method	Mitigating factors	Analytic representation
Rasch difficulty	1.000	0.484* ^a	0.182*	0.253*	0.034	0.168*	-0.295*	-0.004	0.434*	0.089	0.404*	0.228*	0.095	-0.464*	0.471*
Item openness	0.484*	1.000	0.216*	0.283*	0.310*	-0.094	0.017	0.014	0.277*	0.005	0.155*	0.061	0.177*	-0.202*	0.261*
Number of words	0.182*	0.216*	1.000	0.201*	0.345*	0.035	0.089	0.045	0.289*	0.051	0.203*	0.151*	0.314*	-0.168*	-0.076
Divergent thinking	0.253*	0.283*	0.201*	1.000	0.100	-0.018	0.137	-0.073	0.344*	-0.006	0.196*	0.264*	0.293*	-0.186*	-0.076
Graphics in item stem	0.034	0.310*	0.345*	0.100	1.000	0.135	0.120	0.113	0.107	0.124	0.011	0.064	0.175*	-0.104	-0.171*
Intuitive physics (negative)	0.168*	-0.094	0.035	-0.018	0.135	1.000	-0.281*	-0.040	0.135	-0.024	0.195*	0.030	0.065	-0.241*	-0.122
Intuitive physics (positive)	-0.295*	0.017	0.089	0.137	0.120	-0.281*	1.000	-0.075*	-0.057*	-0.132	-0.115	-0.014	0.324*	0.085	-0.209*
Near transfer	-0.004	0.014	0.045	-0.073	0.113	-0.040	-0.075	1.000	-0.473*	8.734*	-0.396*	-0.226*	-0.120	-0.044	0.121
Exploration	0.434*	0.277*	0.289*	0.344*	0.107	0.135	-0.057	-0.473*	1.000	-0.215*	0.760*	0.463*	0.265*	-0.282*	0.049
Relationship	0.089	0.005	0.051	-0.006	0.124	-0.024	-0.132	0.734*	-0.215*	1.000	-0.450*	-0.130	-0.019	-0.021	0.115
Related relationships	0.404*	0.155*	0.203*	0.196*	0.011	0.195*	-0.115	-0.396*	0.760*	-0.450*	1.000	0.509*	0.047	-0.307*	0.121
Visualization	0.228*	0.061	0.151*	0.264*	0.064	0.030	-0.014	-0.226*	0.463*	-0.130	0.509*	1.000	-0.019	-0.255*	-0.081
Experimental method	0.095	0.177*	0.314*	0.293*	0.175*	0.065	0.324*	-0.120	0.265*	-0.019	0.047	-0.019	1.000	0.007	-0.116
Mitigating factors	-0.464*	-0.202*	-0.168*	-0.186*	-0.104	-0.241*	0.085	-0.044	-0.282*	-0.021	-0.307*	-0.255*	0.007	1.000	-0.045
Analytic representation	0.471*	0.261*	-0.076	-0.076	-0.171*	-0.122	-0.209*	0.121	0.049	0.115	0.121	-0.081	-0.116	-0.045	1.000

^aSignificant at the $p < 0.05$ level.

TABLE XIV. Partial correlation coefficients.

	Item openness	Number of words	Divergent thinking	Presence of graphics	Intuitive physics (negative)	Intuitive physics (positive)	Near transfer	Exploration	Relationship	Related relationships	Visualization	Experimental method	Mitigating factors	Analytic representation
Rasch difficulty	0.366	0.053	0.120	-0.105	0.124	-0.239	-0.086	0.010	0.164	0.139	0.055	0.122	-0.364	0.423

- [1] E. F. Redish, *Teaching Physics with the Physics Suite* (Wiley, New York, 2003).
- [2] L. Petrovic, *External Assessment of Student Achievement at Primary School Level, An Expert's Report* (Standards and Assessment Agency for Federation of BiH and RS, Sarajevo, 2006).
- [3] I. V. S. Mullis, M. O. Martin, G. J. Ruddock, C. Y. O'Sullivan, A. Arora, and E. Erberber, TIMSS 2007 Assessment Frameworks, TIMSS & PIRLS International Study Center, Boston College, Chestnut Hill, MA, 2006, <http://timss.bc.edu/TIMSS2007/frameworks.html>.
- [4] J. F. Olson, M. O. Martin, and I. V. S. Mullis, TIMSS 2007 Technical Report, TIMSS & PIRLS International Study Center, Boston College, Chestnut Hill, MA, 2008, <http://timss.bc.edu/TIMSS2007/techreport.html>; M. O. Martin, I. V. S. Mullis, and P. Foy, TIMSS 2007 International Science Report, TIMSS & PIRLS International Study Center, Boston College, Chestnut Hill, MA, 2008, <http://timss.bc.edu/timss2007/sciencereport.html>.
- [5] C. Chalifour and D. E. Powers, The relationship of content characteristics of GRE analytical reasoning items to their difficulties and discriminations, *J. Educ. Measure.* **26**, 120 (1989).
- [6] L. Cohen, L. Manion, and K. Morrison, *Research Methods in Education* (Routledge, New York, 2006).
- [7] C. V. Rosca, Ph.D. thesis, Boston College, 2004.
- [8] F. E. Weinert, *Leistungsmessungen in Schulen* (Beltz Verlag, Weinheim, 2001).
- [9] A. Kauertz, Ph.D. thesis, University Duisburg-Essen, 2007.
- [10] A. Hotiu, M.S. thesis, Florida Atlantic University, 2007.
- [11] TIMSS 2007 International Database, <http://timss.bc.edu/timss2007/idb Ug.html> (2009).
- [12] R. Teodorescu, C. Bennhold, and G. Feldman, in *Proceedings of the Physics Education Research Conference, 2008*, edited by M. Sabella, C. Henderson, and L. Hsu (AIP, Melville, NY, 2008).
- [13] M. McCloskey, Intuitive physics, *Sci. Am.* **248**, 122 (1983).
- [14] A. diSessa, Toward an epistemology of physics, *Cogn. Instr.* **10**, 105 (1993).
- [15] J. Clement, in *Implicit and Explicit Knowledge*, edited by D. Tirosh (Ablex, Hillsdale, NJ, 1994).
- [16] J. P. Guilford, The structure of intellect, *Psychol. Bull.* **53**, 267 (1956).
- [17] J. K. Gilbert, M. Reiner, and M. Nakhleh, *Visualization: Theory and Practice in Science Education* (Springer, Dordrecht, 2008).
- [18] N. Nersessian, *Creating Scientific Concepts* (MIT Press, Cambridge, MA, 2008).
- [19] D. Draxler, Ph.D. thesis, University Duisburg-Essen, 2005.
- [20] I. A. Halloun, *Modeling Theory in Science Education* (Springer, Dordrecht, 2006).
- [21] J. Clement, *Creative Model Construction in Scientists and Students: The Role of Imagery, Analogy, and Mental Simulation* (Springer, Berlin, 2008).
- [22] A. Zietsman and J. Clement, The role of extreme case reasoning in instruction for conceptual change, *J. Learn. Sci.* **6**, 61 (1997).
- [23] S. P. Marshall, in *The Teaching and Assessing of Mathematical Problem Solving*, edited by R. I. Charles and E. A. Silver (Lawrence Erlbaum Associates and the National Council of Teachers of Mathematics, Reston, VA, 1988).
- [24] SAA 2006 Database, Sarajevo office of the Agency for Pre-school, Primary and Secondary Education in BiH, 2006.
- [25] L. M. Wu, J. R. Adams, R. M. Wilson, and S. A. Haldane, *Acer CONQUEST 2.0: Generalised Item Response Modeling Software* (Acer Press, Camberwell, Victoria, 2007).
- [26] M. Planinic, L. Ivanjek, and A. Susac, Rasch model based analysis of the Force Concept Inventory, *Phys. Rev. ST Phys. Educ. Res.* **6**, 010103 (2010).
- [27] B. D. Wright and M. Linacre, Reasonable mean-square fit values, *Rasch Measure. Trans.* **8**, 370 (1994).
- [28] S. Luppescu, Virtual equating, *Rasch Measure. Trans.* **19**, 1025 (2005).
- [29] Winsteps Help for Rasch Analysis, <http://www.winsteps.com/winman/equating.htm>.
- [30] J. Miles and M. Shelvin, *Applying Regression and Correlation: A Guide for Students And Researchers* (SAGE, London, 2001).
- [31] A. Field, *Discovering Statistics using SPSS* (SAGE, London, 2005).
- [32] J. L. Fleiss, Measuring nominal scale agreement among many raters, *Psychol. Bull.* **76**, 378 (1971).
- [33] D. Cook and S. Weisberg, *Residuals and Influence in Regression* (Chapman & Hall, London, 1982).
- [34] H. White, A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity, *Econometrica* **48**, 817 (1980).
- [35] J. R. Landis and G. G. Koch, The measurement of observer agreement for categorical data, *Biometrics* **33**, 159 (1977).
- [36] J. L. Fleiss, *Statistical Methods for Rates and Proportions* (Wiley, New York, 1981).
- [37] J. M. Linacre and B. D. Wright, The length of a logit, *Rasch Measure. Trans.* **3**, 54 (1989).
- [38] T. de Jong and M. Ferguson-Hessler, Types and qualities of knowledge, *Educ. Psychol.* **31**, 105 (1996).
- [39] J. Sweller, J. van Merriënboer, and F. Paas, Cognitive architecture and instructional design, *Educ. Psychol. Rev.* **10**, 251 (1998).
- [40] V. Mesic, in *Proceedings of the International Conference on TIMSS 2007*, edited by N. Suzic and J. Ibrakovic (Agency for Pre-school, Primary and Secondary Education in BiH, Sarajevo, 2010).